jc549 U.S. PTO
08/14/98

Please type a plus sign (+) inside this box → [+]

PTO/SB/05 (12/97)
Approved for use through 09/30/00. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	003239.P010	*Total Pages*	41
First Named Inventor or Application Identifier	Zifei Peter Wang		
Express Mail Label No.	EL105935674US		

jc549 U.S. PTO
09/134212
08/14/98

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. [X] Fee Transmittal Form *(Submit an original, and a duplicate for fee processing)*
2. [X] Specification [*Total Pages* 27]
   (preferred arrangement set forth below)
   - Descriptive title of the Invention
   - Cross References to Related Applications
   - Statement Regarding Fed sponsored R & D
   - Reference to Microfiche Appendix
   - Background of the Invention
   - Brief Summary of the Invention
   - Brief Description of the Drawings (*if filed*)
   - Detailed Description
   - Claim(s)
   - Abstract of the Disclosure
3. [X] Drawing(s) *(35 CFR 113)* [*Total Sheets* 7]
4. Oath or Declaration [*Total Pages* 2]
   a. [X] Newly executed (original copy)
   b. [ ] Copy from a prior application (37 CFR 1.63(d))
   (for continuation/divisional with Box 17 completed)
   [Note Box 5 below]
   i. [ ] DELETION OF INVENTOR(S)
   Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. [ ] Incorporation By Reference *(usable if Box 4b is checked)*
   The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. [ ] Microfiche Computer Program *(Appendix)*
7. Nucleotide and/or Amino Acid Sequence Submission
   (if applicable, all necessary)
   a. [ ] Computer Readable Copy
   b. [ ] Paper Copy (identical to computer copy)
   c. [ ] Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. [X] Assignment Papers (cover sheet & document(s))
9. [ ] 37 CFR 3.73(b) Statement [ ] Power of Attorney
   (when there is an assignee)
10. [ ] English Translation Document *(if applicable)*
11. [ ] Information Disclosure Statement (IDS)/PTO - 1449 [ ] Copies of IDS Citations
12. [ ] Preliminary Amendment
13. [ ] Return Receipt Postcard (MPEP 503)
    (Should be specifically itemized)
14. [ ] Small Entity Statement(s) [ ] Statement filed in prior application, Status still proper and desired
15. [ ] Certified Copy of Priority Document(s)
    (if foreign priority is claimed)
16. [ ] Other: . . . . . .

17. **If a CONTINUING APPLICATION,** *check appropriate box and supply the requisite information:*
    [ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No:____/_____

18. CORRESPONDENCE ADDRESS

[ ] *Customer Number of Bar Code Label* *(Insert Customer No. or Attach bare code label here)* *or* [X] *Correspondence address below*

NAME	BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP				
ADDRESS	12400 Wilshire Boulevard, Seventh Floor				
CITY	Los Angeles	*STATE*	California	*ZIP CODE*	90025
COUNTRY	U.S.A.	*TELEPHONE*	(714) 557-3800	*FAX*	(714) 557-3347

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

PTO/SB/17 (10/97)
Approved for use through 09/30/00. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

FEE TRANSMITTAL

Note: Effective October 1, 1997.
Patent fees are subject to annual revision.

TOTAL AMOUNT OF PAYMENT ($) 912.00

Complete If Known	
Application Number	
Filing Date	August 14, 1998
First Named Inventor	Zifei Peter Wang
Group Art Unit	
Examiner Name	
Attorney Docket Number	003239.P010

METHOD OF PAYMENT (check one)

1. ☒ The Commissioner is hereby authorized to charge indicated fees and credit any over payments to:

Deposit Account Number: 02-2666

Deposit Account Name: Blakely, Sokoloff, Taylor & Zafman LLP

☒ Charge Any Additional Fee Required Under 37 CFR 1.16 and 1.17 ☐ Charge the Issue Fee Set in 37 CFR 1.18 at the Mailing of the Notice of Allowance, 37 CFR 1.311(b)

2. ☒ Payment Enclosed:
☒ Check ☐ Money Order ☐ Other

FEE CALCULATION (fees effective 10/01/96)

1. FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
101	790	201	395	Utility filing fee	$790
106	330	206	165	Design filing fee	
107	540	207	270	Plant filing fee	
108	790	208	395	Reissue filing fee	
114	150	214	75	Provisional filing fee	

SUBTOTAL (1) ($) 790.00

2. CLAIMS

		Extra		Fee from below		Fee Paid
Total Claims	18	-20 =	0	X $22.00	=	0.00
Independent Claims	4	-3 =	1	X $82.00	=	82.00
Multiple Dependent Claims					=	

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
103	22	203	11	Claims in excess of 20
102	82	202	41	Independent claims in excess of 3
104	270	204	135	Multiple Dependent claim
109	82	209	41	Reissue independent claims over original patent
110	22	210	11	Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($) 82.00

FEE CALCULATION (continued)

3. ADDITIONAL FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
105	130	205	65	Surcharge - late filing fee or oath	
127	50	227	25	Surcharge - late provisional filing fee or cover sheet.	
139	130	139	130	Non-English specification	
147	2,520	147	2,520	For filing a request for reexamination	
112	920*	112	920*	Requesting publication of SIR prior to Examiner action	
113	1,840*	113	1,840*	Requesting publication of SIR after Examiner action	
115	110	215	55	Extension for response within first month	
116	400	216	200	Extension for response within second month	
117	950	217	475	Extension for response within third month	
118	1,510	218	755	Extension for response within fourth month	
119	310	219	155	Notice of Appeal	
120	310	220	155	Filing a brief in support of an appeal	
121	270	221	135	Request for oral hearing	
138	1,510	138	1,510	Petition to institute a public use proceeding	
140	110	240	55	Petition to revive unavoidably abandoned application	
141	1,320	241	660	Petition to revive unintentionally abandoned application	
142	1,320	242	660	Utility issue fee (or reissue)	
143	450	243	225	Design issue fee	
144	670	244	335	Plant issue fee	
122	130	122	130	Petitions to the Commissioner	
123	50	123	50	Petitions related to provisional applications	
126	240	126	240	Submission of Information Disclosure Stmt	
581	40	581	40	Recording each patent assignment per property (times number of properties)	40.00
146	790	246	395	Filing a submission after final rejection (37 CFR 1.129(a))	
149	790	249	395	For each additional invention to be examined (37 CFR 1.129(b))	
Other fee (specify)					
Other fee (specify)					

* Reduced by Basic Filing Fee Paid

SUBTOTAL (2) ($) 40.00

SUBMITTED BY				Complete (if applicable)	
Typed or Printed Name	William W. Schaal, Reg. No. 39, 018			Reg. Number	
Signature		Date	08/14/98	Deposit Account User ID	02-2666

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

Atty. Docket No.: 003239.P010
Express Mail No.: EL105935674US

APPLICATION FOR UNITED STATES PATENT

FOR

A VOICE ACTIVITY DETECTOR FOR PACKET VOICE NETWORK

Inventor:

ZIFEI PETER WANG

Prepared by:

BLAKELY SOKOLOFF TAYLOR & ZAFMAN
12400 Wilshire Boulevard, Seventh Floor
Los Angeles, California 90025-1026
(714) 557-3800

BACKGROUND

1. Field

The present invention relates to the field of data communications. In particular, this invention relates to a system and method for enhancing the reliability of voice activity detection.

2. General Background

For many years, discontinuous transmission (DTX) systems have been installed to conserve bandwidth over packet voice/data networks. Bandwidth conservation is accomplished by detecting when a caller is speaking and transmitting speech packets generated by a speech coder during those periods of time. For the remaining periods of time when the caller is not speaking, certain DTX systems have been configured to transmit a background noise level tracked by a voice activating detector. This background noise level is subsequently used to replicate the background silence gaps between communications, which are a considerable portion of normal speech communications.

Conventional DTX systems consist of a voice activity detector (VAD) and a comfort noise generator (CNG). Normally, a "voice activity detector" (VAD) is software processed by circuitry to digitize an analog signal (e.g., voice and/or background noise) and to determine whether or not a particular segment of the digitized analog signal represents a person's voice. Since the range of a person's voice is dynamic, in some situations varying 20-40 decibels (dB), and background noise can vary moment to moment, a number of

different parameters have been used by conventional VADs to discern voice activity.

For example, an IEEE publication entitled "Application of an LPC distance measure to the voice-unvoiced-silence detection problem," authored by L.R. Rabiner and M.R. Sambur, describes a voice activity detector (VAD) performing a pattern recognition approach on incoming digitally sampled signals to detect voice activity. In particular, this VAD creates templates of parameters for voiced, unvoiced (e.g., tailing off sounds for certain words) and silence segments of speech. Each template includes five parameters: the energy of the signal (E_S); the zero-crossing rate of the signal (N_Z); the autocorrelation coefficient at unit sample delay (C_1); the first order predictor coefficient (A_1); and the normalized prediction error (E_p). Through probability calculations, decision logic compares the templates with a sampled segment of an incoming signal to determine whether the segment represents voice, unvoice or silence. The disadvantage associated with this VAD is that it is extremely difficult to find a set of reliable templates to distinguish between a variety of speech signals and numerous levels of background noise found in different environments.

Another example of VAD involves the use of linear prediction coefficients (LPC) which are calculated in the speech coder. While taking advantage of the LPCs calculated in the speech coders reduce computational power consumption by the VAD, it also has encountered a number of disadvantages. For example, speech coders in accordance with the International Telegraph and Telephone Consultive Committee (CCITT) G.729B standards perform linear predictive coding differently than speech coders in accordance with CCITT G.723 standards. As a result, there does not

exist a VAD which can be used by virtually all types of speech coders. Instead, depending on the type of speech coder implemented, the VAD must be modified to operate in combination with that speech coder. This increases overall ownership costs and the difficulty in upgrading the DTX system.

Over the last few years, MICOM Communications Corporation of Simi Valley, California, has produced voice/data networking products for DTX systems that utilize a universal energy-based VAD. The voice/data networking products includes a dual-mode speech coding function in order to achieve bandwidth efficiency. In a VOICE mode, a selected speech coder is responsible for compressing voice signals before transmission and for decompressing the voice signals upon reception. In a SILENCE SUPPRESSION mode, only the background noise level signal is transmitted, from which white noise is regenerated at the destination.

Currently, two parameters are used by this universal VAD function in order to determine whether the voice/data networking product is operating in a VOICE mode or a SILENCE SUPPRESSION mode. These parameters include (i) short-term tracking energy and (ii) long-term tracking energy. The "short-term tracking energy" is an accumulation of signal energy associated with voice signaling and background noise level, and thus, is represented by equation (1).

$$(1) \qquad E_{trk}(k) = \alpha \times E_{db}(k) + (1-\alpha) \times E_{trk}(k-1),$$

$$\text{where } \alpha = \begin{cases} 1/4 \text{ if } E_{db}(k) \geq E_{trk}(k-1), \text{ or} \\ 1/8 \text{ otherwise.} \end{cases}$$

$E_{dB}(k)$ denotes the current frame energy in decibels and is equivalent to the following: $10 \log_{10} \left(\sum_{n=0}^{n=N-1} s(n)^2 \right)$ where "N" represents the number of samples per frame.

$E_{trk}(k-1)$ denotes the short-term tracking energy for the previous frame.

The "long-term tracking energy" represents the background noise level associated with incoming audio and is measured by equation (2).

(2) $E_l(k) = \min\{\beta E_l(k-1) + (1-\beta)E_s(k), E_{max}\}$, where

β=0.875; and

E_{max} denotes the maximum background level.

As a result, when the calculated value of the long-term tracking energy approaches the calculated value of the short-term tracking energy, the VAD predicts that a segment of sampled signals associated with a current frame is likely to be silence. One problem that has been encountered is that this conventional VAD is subject to increased switching between VOICE mode and SILENCE SUPPRESSION mode during long periods of silence, where the long-term tracking energy naturally approaches the short-term tracking energy. This increasing switching, referred to as "in/out effects," causes audio volume fluctuations detectable by the human ear.

Hence, it would be advantageous to provide a system and method for enhancing reliability of voice activity detection through development of an improved, universal VAD which relies on a peak-to-mean likelihood ratio. The peak-to-mean likelihood ratio reduces the occurrence of the in/out effects by further assisting the VAD, in certain instances, to determine whether an incoming analog signal represents voice or silence.

SUMMARY OF THE INVENTION

The present invention relates to a voice activity detector, being either software executable by a processing unit or firmware, which predicts whether an audio frame represents a voice signal or silence. This prediction is based the analysis of a number of parameters, including a short-term averaged energy (STAE), a long-term averaged energy (LTAE), and a peak-to-mean likelihood ratio (PMLR).

In one embodiment, to predict whether a frame represents voice or silence, an initial determination is made whether a sum of the STAE and a factor is greater than the LTAE. If the sum is less than the LTAE, the audio frame represents silence. Otherwise, a second determination is made as to whether the difference between the LTAE and the STAE is less than a predetermined threshold. In the event that the difference between the LTAE and the STAE is less than the predetermined threshold, the PMLR is determined and compared to a selected threshold. If the PMLR is greater than the selected threshold, the audio frame represents a voice signal. Otherwise, it represents silence.

Select Exhibit

BRIEF DESCRIPTION OF THE DRAWINGS

The features and advantages of the present invention will become apparent from the following detailed description of the present invention in which:

Figure 1 is an illustrative diagram of a system comprising a first networking device operating in accordance with the present invention.

Figure 2 is an illustrative diagram of an embodiment of a communication module employed within the first networking device of Figure 1.

Figure 3 is an illustrative flowchart of the operations of the first networking device of Figure 1.

Figure 4 is an illustrative block diagram of the data structure of a service frame.

Figure 5 is an illustrative block diagram of the data structure of a silence suppression frame.

Figure 6 is an illustrative flowchart of the operations of the second networking device.

Figure 7 is an illustrative block diagram of the operations of the comfort noise generator.

Figure 8 is an illustrative flowchart of the operations of the voice activating detector.

Figure 9 is an illustrative block diagram of hardware for calculating the average peak-mean ratio.

Figure 10 is an illustrative block diagram of a state diagram of a decision smoothing state machine for further reduction of in/out effects.

DELPHION

DETAILED DESCRIPTION OF AN EMBODIMENT

Herein, embodiments of the present invention relates to a system and method for enhancing reliability in voice activity detection. This is accomplished by an improved voice activity detector in which an additional parameter, a peak-to-mean likelihood ratio (PMLR), is used in combination with long-term averaged energy and short-term averaged energy parameters to determine whether various segments of audio constitute voice or silence. The use of the peak-to-mean likelihood ratio by the voice activity detector will reduce audio degradation currently experienced by conventional DTX systems.

Herein, certain terminology is used to describe various features of the present invention. In general, a "system" comprises one or more networking devices coupled together through corresponding signal lines. A "networking device" comprises a digital platform such as, for example, a MARATHON™ frame relay product by Nortel/MICOM, a voice-over Asynchronous Transfer Mode (ATM) product such as Passport 4740™ by Nortel/MICOM, cellular telephones operating in accordance with a cellular communication standard (e.g., GSM) and the like. Such a digital platform usually comprises software and/or hardware to perform analog to linear conversion, echo cancellation, speed coding, etc.. A "signal line" includes any communications link capable of transmitting digital information at some ascertainable bandwidth. Examples of a signal line include a variety of mediums such as T1/E1, frame relay, private leased line, satellite, microwave, fiber optic, cable, wireless communications (e.g., radio frequency "RF") or even a logical link.

Additionally, "information" generally comprises a signal having one or more bits of data, address, control or any combination thereof. A "communication module" includes a voice activity detector used to determine whether various segments of audio constitute voice or silence. In this embodiment, the "voice activity detector" (VAD) is software; however, it is contemplated that the VAD may be implemented in its entirety as hardware or firmware being a combination of hardware and software.

Referring to Figure 1, an illustrative embodiment of a system utilizing the present invention is shown. Herein, system 100 includes a first networking device (source) 110 coupled to a second networking device (destination) 120 via a signal line 130. Herein, networking device 110 receives analog audio signals 140 as input and digitizes the audio to produce pulse code modulation (PCM) audio for example. The PCM audio is separated into multiple frames, where various signal characteristics of each frame are analyzed by a voice activating detector (VAD) as described below in Figure 8. From these signal characteristics, first networking device 110 can determine whether to transmit a compressed audio frame (referred to as a "service frame") or to transmit a silence suppression frame providing a noise background level as described below.

Referring now to Figure 2, first networking device 110 comprises a communication module 200. Communication module 200 includes a substrate 210 which is formed with any type of material or combination of materials upon which integrated circuit (IC) devices can be attached. Communication module 200 is adapted to a connector 220 in order to exchange information with other logic mounted on a circuit board 260 of networking device 110 for example. Any style for connector 220 may be used,

including a standard female edge connector, a pin field connector, a socket, a network interface card (NIC) connection and the like.

As shown, communication module 200 includes memory 230 and a processing unit 240. In this embodiment, memory 230 includes off-chip volatile memory to contain software which, when executed by processing unit 240, performs voice activity detection. Of course, non-volatile memory may be used in combination with or in lieu of volatile memory. Processing unit 240 includes, but is not limited or restricted to a general purpose microprocessor, a digital signal processor, a micro-controller or any other logic having software processing capabilities. Processing unit 240 includes on-chip internal memory (M) 250 to receive information from memory 230 for internal storage thereby enhancing its processing speed.

Referring now to Figure 3, an illustrative flowchart of the operations performed by first networking device 110 is shown. Initially, first networking device 110 receives analog audio and digitizes the audio. For this example, the audio may be converted into PCM audio (block 300). The PCM audio is modified by an echo canceler (block 310), in order to eliminate echo returned from second networking device 120 of Figure 1, and thereafter, each frame of the PCM audio is analyzed by a voice activity detector (VAD). For example, the VAD may be software executed by processing unit 240 of Figure 2 (block 320). Based on signal characteristics of each PCM audio frame, a determination is made whether the frame constitutes voice or silence (block 330).

If the frame is determined to be voice, first networking device 110 enters into a VOICE mode. In this mode, the PCM audio frame is loaded into

a speech coder which compresses the PCM audio frame to produce a service frame as shown in Figure 4 (block 340). The service frame 260 includes a header 265 to identify the frame and payload 270 to contain compressed audio. Such compression is performed in accordance with any existing or later developed compression function.

Alternatively, if the frame is determined to be silence, first networking device enters into a SILENCE SUPPRESSION mode. In this mode, a silence suppression frame (see Figure 5) is transmitted to the second networking device (block 350). The silence suppression frame 275 comprises a header 280, a first field 285 to contain a background noise level being an energy value representing the background noise, and a second field 290 to contain the complement of the background noise level. The complement is included for error checking. This process, inclusive of voice activity detection, continues for each PCM audio frame (block 360).

Referring now to Figure 6, an illustrative flowchart of the operations performed by second networking device 120 of Figure 1 is shown. Upon receiving a frame of information (block 400), second networking device 120 determines whether a silence suppression frame has been received (block 410). If so, the background noise level recovered from the silence suppression frame is loaded into a comfort noise generator (CNG). The CNG produces comfort noise samples based on the received background level in order to avoid audio artifacts such as in-out effects (block 420).

In particular, as shown in Figure 7, CNG 500 includes linear factor calculator 510 to handle various ranges of background noise levels. Each of these ranges (in dB) is mapped into a linear factor 520 which is used to scale a

constant level of noise 530 supplied by a random number generator. The scaled white noise 540 is then passed through a first order 1/f filter 550 to obtain the pink noise samples. The resultant pink noise is a regeneration of the background noise at the source. Thereafter, the pink noise samples are placed in an analog format (block 430) as shown in Figure 6.

Referring still to Figure 6, in the alternative event that a service frame is detected so no error condition is triggered (blocks 440-450), the service frame is transferred to a speech decoder to recover a substantial portion of the original PCM audio (block 460). Thereafter, the PCM audio is placed in an analog format (block 430).

Referring to Figure 8, an illustrative flowchart of the operations of the voice activity detector (VAD) is shown. Initially, each audio frame is collected for N samples per frame (block 600). In this embodiment, the sampling number "N" is approximately 80 samples per frame, but may be any number of samples up to the size supported by a speech coder. After the audio frame has been collected, a number of signal parameters are calculated, including the short-term averaged energy, the long-term averaged energy, and the peak-to-mean likelihood ratio.

Before calculating the short-term averaged energy and the long-term averaged energy, the energy associated with the current audio frame is calculated (block 610). This is accomplished by squaring each voice sample (s_i) for the current audio frame and summing the squared result. The frame energy is defined by equation (3).

$$(3) \qquad E = \sum_{i=0}^{N-1} (s_i)^2$$

After the current frame energy has been calculated, it is converted into a decibel (dB) value (block 620). This provides a larger dynamic range to handle a greater energy variance for each sampled audio frame. The frame energy (in dB) is calculated as shown in equation (4).

(4) $E_{dB} = 10 \log_{10}(E)$

After calculating E_{dB} for the current frame, the short term averaged energy may be calculated (block 630). The short-term averaged energy (STAE) is an accumulation of signal energy associated with successive PCM audio frames. The current frame energy E_{dB} and the STAE for the previous frame are weighted by predetermined factors "α" and "1-α" so that the resultant value is the STAE for the current frame. The selection of the factor "α" may be set through simulations. Herein, the STAE is defined in equation (5) as:

(5) $E_s(k) = \alpha \times E_{dB}(k) + (1-\alpha) \times E_s(k-1)$, where

$$\alpha = \begin{cases} 0.125 \text{ if } E_{dB}(k) \geq E_s(k-1) \\ 0.25 \text{ otherwise.} \end{cases}$$

"α" denotes a selected factor of the energy of a current PCM audio frame to be added to the accumulated average.

"$E_{dB}(k)$" denotes the current frame energy in decibels; and

"$E_s(k-1)$" denotes the prior short-term averaged energy value.

Along with the STAE, the "long-term averaged energy" (LTAE) is calculated (block 640). The LTAE is defined as an additional level of accumulation to track the background noise level and, for this embodiment, is updated in accordance with equation (6):

$$(6)\qquad E_x(k)=\begin{cases}\min\{\beta E_x(k-1)+(1-\beta)E_s(k),\ E_{max}\}, & \text{if } E_x(k-1)>E_s(k)\\ \min\{E_x(k-1)+\delta E_x,\ E_{max}\}, & \text{otherwise}\end{cases}$$

where $\beta = 0.875$

$$\delta E_x=\begin{cases}1 & \text{if previous form is voice,}\\ \dfrac{1}{16} & \text{otherwise.}\end{cases}$$

E_{max} denotes the maximum background level being set to -30dBm0.

In the case where $E_x(k\text{-}1)<E_s(k)$, instead of adaptively updating LTAE, we apply a jump (δE_x). By doing so, we can update the LTAE promptly when there is a sudden change in background noise level.

Next, a peak-to-mean ratio (PMR) is calculated in order to determine the peak-to-mean likelihood ratio (block 650). The PMR comprises a ratio between the absolute value of a maximum sampled signal and the summation of the values for all (N) sampled signals for the current frame as shown in equation (7). Therefore, as the value of the PMR increases, there is a greater likelihood that the current frame represents silence because a waveform associated with silence has lesser energy than a waveform associated with voice.

$$(7)\qquad PMR = \frac{\max\{|s_i|\}}{\sum_{i=0}^{N-1}|s_i|}$$

After the PMR is calculated, an average peak-to-mean ratio (APMR) is now determined (block 660) for use in calculating the peak-mean likelihood ratio (PMLR). The reason for calculating APMR is to prevent frequent

alterations between VOICE mode and SILENCE SUPPRESSION mode based on environmental conditions (e.g., speaker talks loudly, noisy environment, etc.). Consequently, the occurrence of an in/out effect is substantially mitigated.

As shown in Figure 9, one technique to calculate the APMR is to implement a circular buffer 700 having depth "M". During analysis by the VAD, the PMR for that frame is inserted into buffer 700. After each insertion, the APMR is calculated by averaging all of the PMRs loaded into buffer 700 based on equation (8):

$$(8) \qquad APMR = \frac{1}{M}\sum_{i=0}^{M-1} PMR_i$$

Referring back to Figure 8, it is contemplated that the PMR and APMR may be used for voice activity detection. The behavior of PMR or APMR may vary, depending on the audible level of the speaker's voice or the background noise. Thus, in this embodiment, a normalized parameter, namely a peak-mean likelihood ratio, is calculated and subsequently used to determine whether a sampled frame represents voice or silence (block 670).

More specifically, the peak-mean likelihood ratio (PMLR) is a parameter which is compared with a predetermined threshold value to determine whether a sampled frame represents voice or silence. This threshold value is programmed during simulation, allowing a customer to select an acceptable tradeoff between voice quality and bandwidth savings.

As shown in equation (9) below, the PMLR is normalized to substantially mitigate modification caused by different speakers and different background noise levels. As a result, PMLR has minimal variation between

audio frames in order to discourage in/out effects due to frequent switching between VOICE mode and SILENCE SUPPRESSION mode. Also, PMLR is independent of frame size, and thus, can operate with speech coders supporting different frame sizes.

To determine the PMLR, the VAD keeps track of the maximum APMR ($APMR_{max}$) and the minimum APMR ($APMR_{min}$) contained in buffer 700 of Figure 9. The contents of buffer 700 may be periodically cleared after a selected period of time has expired or after a selected number (S) of calls ($S \geq 1$). From these values and the APMR associated with the current audio frame, the PMLR can be measured by equation (9).

$$(9) \qquad PMLR_k = \frac{(APMR_{max} - APMR_k)}{(APMR_{max} - APMR_{min})}$$

In block 680, based on the STAE, LTAE and PMLR parameters, the VAD performs a bifurcated decision process to determine whether a sampled audio frame is voice or silence. A first determination is whether the combination of the STAE and a selected factor is greater than the LTAE as shown in equation (10). The factor is set based on simulation results, which was determined to be 2 dB in this embodiment. Of course, as the factor is increased, more bandwidth will be conserved because there is greater probability for the system to be placed in a VOICE mode.

$$(10) \qquad STAE + factor(2dB) \stackrel{?}{\geq} LTAE$$

If the combination is greater than the LTAE, the sampled audio frame is initially considered to be voice. As a result, the VAD performs a second determination. This determination involves ascertaining the PMLR when the LTAE and the STAE differ by less than a predetermined threshold. The

predetermined threshold is determined to be 4 dB in this embodiment. In mathematical terms:

$$|LTAE - STAE| < \text{Threshold (4dB)}$$

When this condition is met, the VAD determines whether the PMLR is less than a selected threshold. The selected threshold is determined to be 0.50 in this embodiment. If the PMLR is less than the selected threshold, the sampled audio frame represents silence. Otherwise, it represents voice. Consequently, the PMLR provides a secondary determination when the LTAE is approaching the STAE to avoid needless in/out effects.

Once the determination has been made that the sampled audio frame is voice or silence, the VAD performs a decision smoothing process (block 690). The decision smoothing function delays the system from switching from the VOICE mode to the SILENCE SUPPRESSION mode immediately after the current frame is detected to be silence. This avoids speech clipping at the end of an utterance.

Referring now to Figure 10, a state diagram concerning the operations of a decision smoothing state machine 800 of the VAD is shown. State machine 800 comprises a VOICE (mode) state 810, a SILENCE SUPPRESSION state 820 and a HANGOVER state 830. For each sampled audio frame, state machine 800 determines the operating state of the system. In the HANGOVER state 830, the system operates as in the VOICE state.

As shown, state machine 800 enters or remains in VOICE state 810 if the current audio frame is determined to be voice as represented by arrows 840, 845 and 850. However, when the current audio frame is determined to

be silence, the operating mode of the system depends on the current state of state machine 800. For example, if state machine 800 is in SILENCE SUPPRESSION state 820, state machine 800 remains in that state as represented by arrow 855. However, if state machine 800 is in VOICE state 810 and the current audio frame is determined to be silence, state machine enters into HANGOVER state 830 as represented by arrow 860. Consequently, only after a predetermined number (Q) of subsequent audio frames are determined to be silence (# of frames $\geq$ Q), state machine 800 enters into SILENCE SUPPRESSION state 820 as represented by arrow 865. However, if prior to that time, the sampled audio frame is determined to be voice, state machine enters into VOICE state 810 as represented by arrow 850. As a result of these operations, speech clipping is substantially avoided.

While certain exemplary embodiments have been described and shown in the accompanying drawings, it is to be understood that such embodiments are merely illustrative of and not restrictive on the broad invention, and that this invention not be limited to the specific constructions and arrangements shown and described, since various other modifications may occur to those ordinarily skilled in the art.

CLAIMS

What is claimed is:

1. A method for enhancing voice activity detection comprising:

determining a peak-to-mean likelihood ratio; and

comparing the peak-to-mean likelihood ratio to a selected threshold to determine whether a current audio frame represents a voice signal.

2. The method of claim 1, wherein prior to determining the peak-to-mean likelihood ratio, the method further comprises:

determining a short-term averaged energy for the current audio frame; and

determining a long-term averaged energy for the current audio frame.

3. The method of claim 2, wherein after determining the short-term averaged energy and the long-term averaged energy, the method further comprises:

determining whether a sum of the short-term averaged energy and a factor is greater than the long-term averaged energy; and

determining that the current audio frame represents silence if the sum is less than the long-term averaged energy, without necessitating a determination of the peak-to-mean likelihood ratio.

4. The method of claim 3, upon determining that the sum is greater than the long-term averaged energy and before determining the peak-to-mean likelihood ratio, the method further comprises:

determining whether a difference between the long-term averaged energy and the short-term averaged energy is less than a predetermined threshold;

determining that the current audio frame represents voice if the difference is greater than the predetermined threshold; and

continuing by determining the peak-to-mean likelihood ratio if the difference is less than the predetermined threshold.

5. The method of claim 2, wherein the determining of the short-term averaged energy comprises:

determining an energy, in decibels, of the current audio frame;

determining a short-term averaged energy for a prior audio frame; and

conducting a weighted average of the energy of the current audio frame and the short-term averaged energy for the prior audio frame.

6. The method of claim 1, wherein the determining a peak-to-mean likelihood ratio comprises

calculating an averaged peak-to-mean ratio for the current audio frame;

determining a maximum averaged peak-to-mean ratio;

determining a minimum averaged peak-to-mean ratio;

determining a first result being a difference between the maximum averaged peak-to-mean ratio and the averaged peak-to-mean ratio for the current audio frame;

determining a second result being a difference between the maximum averaged peak-to-mean ratio and the minimum averaged peak-to-mean ratio; and

conducting a ratio between the first result and the second result to produce the peak-to-mean likelihood ratio.

7. A communication module comprising:

a substrate;

a processing unit placed on the substrate; and

a memory coupled to the processing unit, the memory to contain a voice activity detector which, when executed by the processing unit, analyzes a short-term averaged energy, a long-term averaged energy, and a peak-to-mean likelihood ratio in order to determine whether a current audio frame represents voice or silence.

8. The communication module of claim 7, wherein the voice activity detector, when executed, controls the processing unit to determine whether a sum of the short-term averaged energy and a predetermined factor is greater than the long-term averaged energy, and to signal that the current audio frame represents silence if the sum is less than the long-term averaged energy.

9. The communication module of claim 8, wherein the voice activity detector, when executed, controls the processing unit to determine

whether a difference between the long-term averaged energy and the short-term averaged energy is less than a predetermined threshold, and to signal that the current audio frame represents voice if the difference is greater than the predetermined threshold.

10. The communication module of claim 9, wherein the voice activity detector, when executed, controls the processing unit to determine the peak-to-mean likelihood ratio, and to compare the peak-to-mean likelihood ratio to a selected threshold to determine whether a current audio frame represents a voice signal.

11. The communication module of claim 10, wherein the voice activity detector, when executed, controls the processing unit to determine a peak-to-mean ratio by (i) sampling an analog signal a predetermined number of times to produce a plurality of sampled signals each having a sampled value, (ii) determining a maximum value of the plurality of sampled signals, and (iii) conducting a ratio between an absolute value of the maximum value and a summation of the sampled values for the plurality of sampled signals.

12. The communication module of claim 10, wherein the voice activity detector, when executed, controls the processing unit to determine an averaged peak-to-mean ratio for the current audio frame by (i) monitoring a maximum averaged peak-to-mean ratio and a minimum averaged peak-to-mean ratio, (ii) determining a first result being a difference between the maximum averaged peak-to-mean ratio and the averaged peak-to-mean ratio for the current audio frame, (iii) determining a second result being a difference between the maximum averaged peak-to-mean ratio and the

minimum averaged peak-to-mean ratio, and (iv) conducting a ratio between the first result and the second result to produce the peak-to-mean likelihood ratio.

13. A machine readable medium having embodied thereon a computer program for processing by a machine, the computer program comprising:

a first routine for determining a peak-to-mean likelihood ratio; and

a second routine for comparing the peak-to-mean likelihood ratio to a selected threshold to determine whether an audio frame being transmitted represents a voice signal.

14. The machine readable medium of claim 13, wherein the computer program further comprising:

a third routine for determining a short-term averaged energy for the audio frame, the third routine being executed before the first and second routines; and

a fourth routine for determining a long-term averaged energy for the audio frame, the fourth routine being executed before the first and second routines.

15. The machine readable medium of claim 14, wherein the computer program further comprising:

a fifth routine for determining whether a sum of the short-term averaged energy and a predetermined factor is greater than the long-term averaged energy, the fifth routine being executed before the first and second routines; and

a sixth routine for determining whether a difference between the long-term averaged energy and the short-term averaged energy is less than a predetermined threshold, the sixth routine being executed after determining that the sum is greater than the long-term averaged energy and before execution of the first and second routines.

16. The machine readable medium of claim 15, wherein the fifth routine determining that the current audio frame represents silence if the sum is less than the long-term averaged energy.

17. The machine readable medium of claim 15, wherein the sixth routine determining that the current audio frame represents voice if the difference is greater than the predetermined threshold.

18. A voice activity detector comprising:

circuitry to determine a short-term averaged energy for an audio frame;

circuitry to determine a long-term averaged energy for the audio frame;

circuitry to determine whether the short-term averaged energy is greater than the long-term averaged energy by a predetermined factor;

circuitry to determine whether a difference between the long-term averaged energy and the short-term averaged energy is less than a predetermined threshold when the short-term averaged energy is greater than the long-term averaged energy by the predetermined factor;

circuitry to determine a peak-to-mean likelihood ratio when the difference between the long-term averaged energy and the short-term averaged energy is less than the predetermined threshold; and

circuitry to comparing the peak-to-mean likelihood ratio to a selected threshold and to determine that the audio frame represents a voice signal when the peak-to-mean likelihood ratio is greater than a selected threshold.

Docket No. 003239.P010

ABSTRACT OF THE DISCLOSURE

A voice activity detector to analyze a short-term averaged energy (STAE), a long-term averaged energy (LTAE), and a peak-to-mean likelihood ratio (PMLR) in order to determine whether a current audio frame being transmitted represents voice or silence. This is accomplished by determining whether a sum of the STAE and a factor is greater than the LTAE. If not, the current audio frame represents silence. If so, a second set of determinations is performed. Herein, a determination is made as to whether the difference between the LTAE and the STAE is less than a predetermined threshold. If so, the current audio frame represents voice. Otherwise, the PMLR is determined and compared to a selected threshold. If the PMLR is greater than the selected threshold, the current audio frame represents a voice signal. Otherwise, it represents silence.

DOCKET NO. 003239.P010

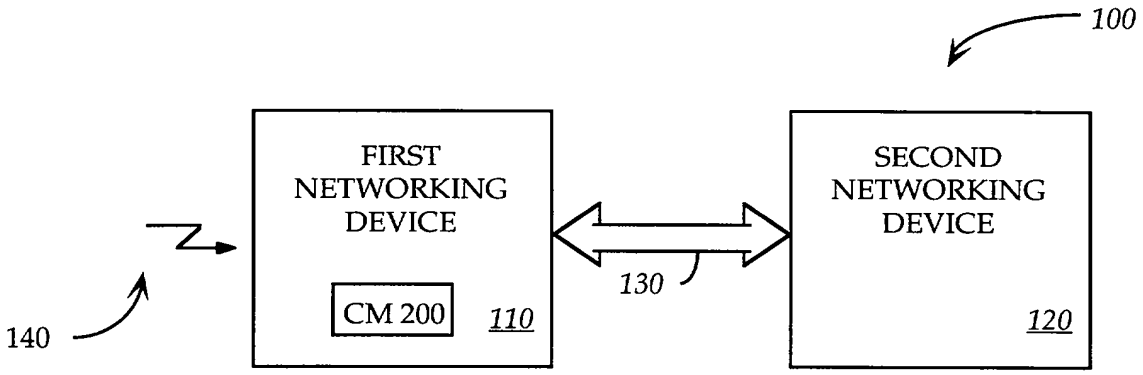

Figure 1

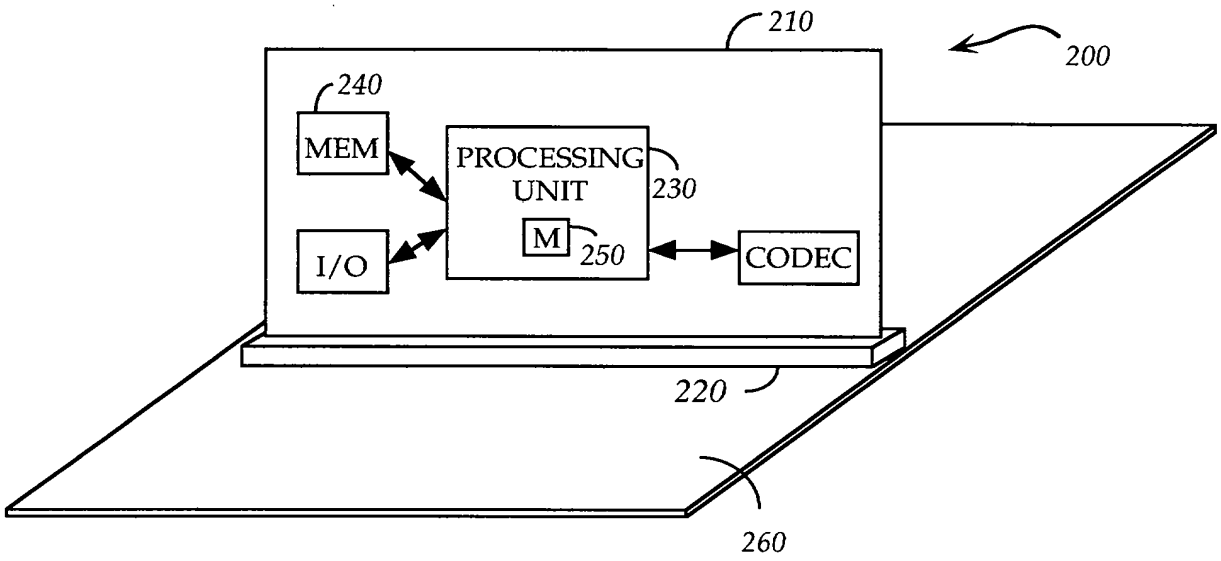

Figure 2

Attorney Docket No. 2004-0168

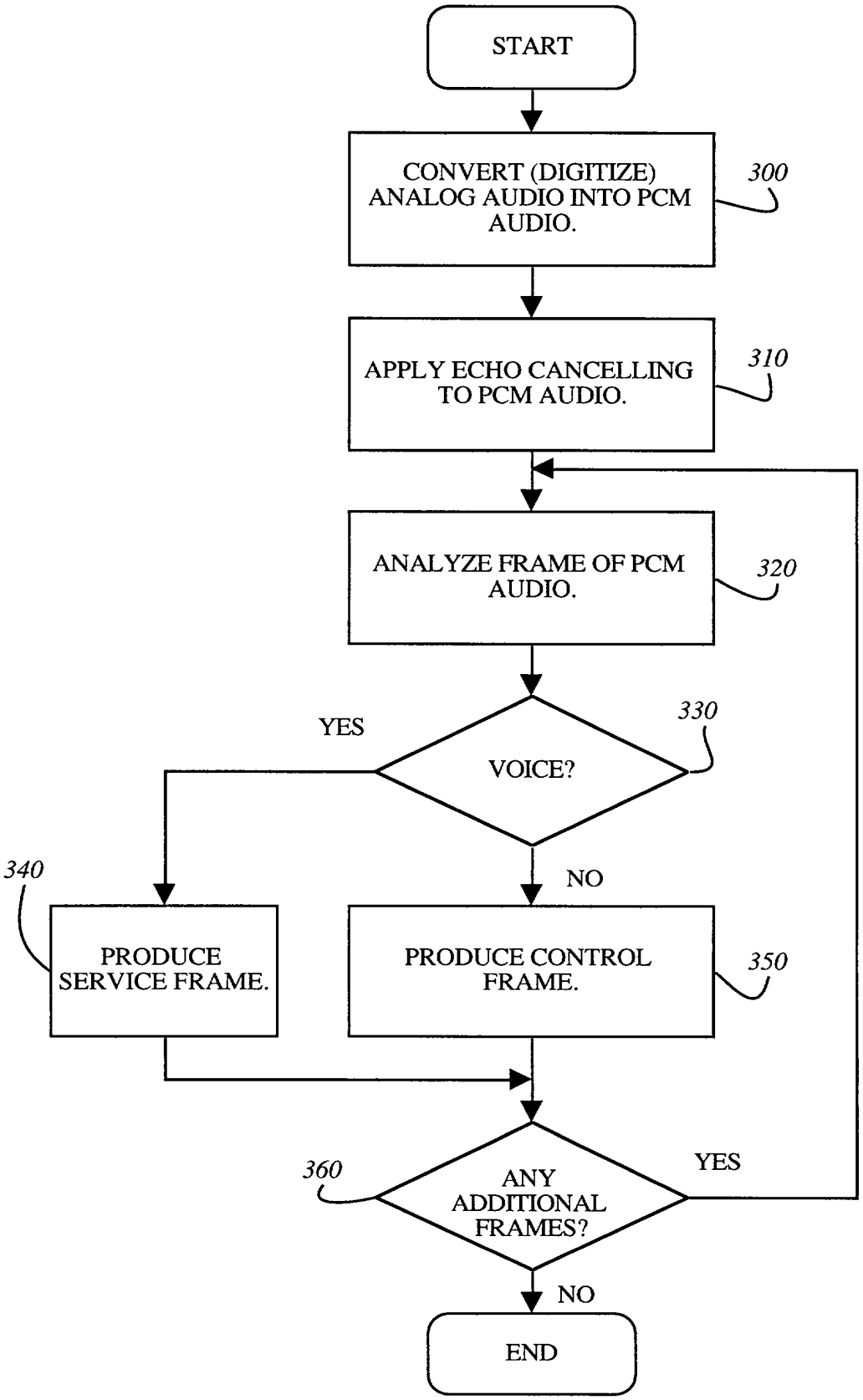

Figure 3

Docket No. 50277-0009

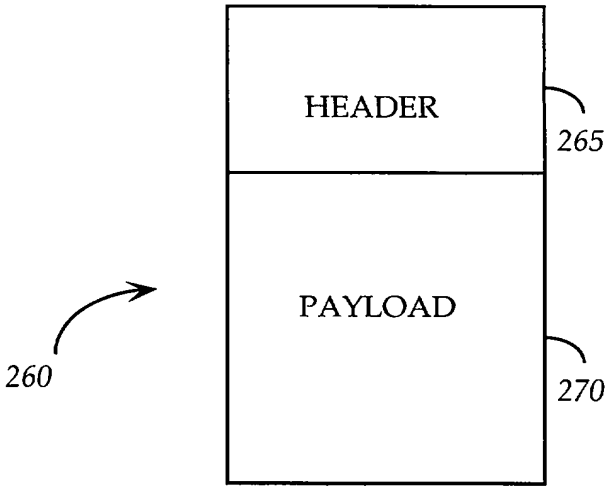

Figure 4

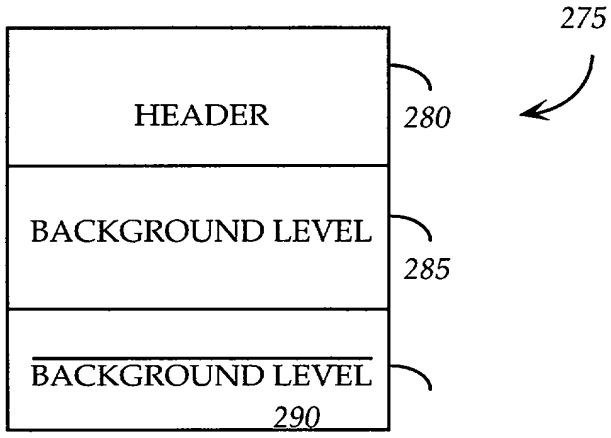

Figure 5

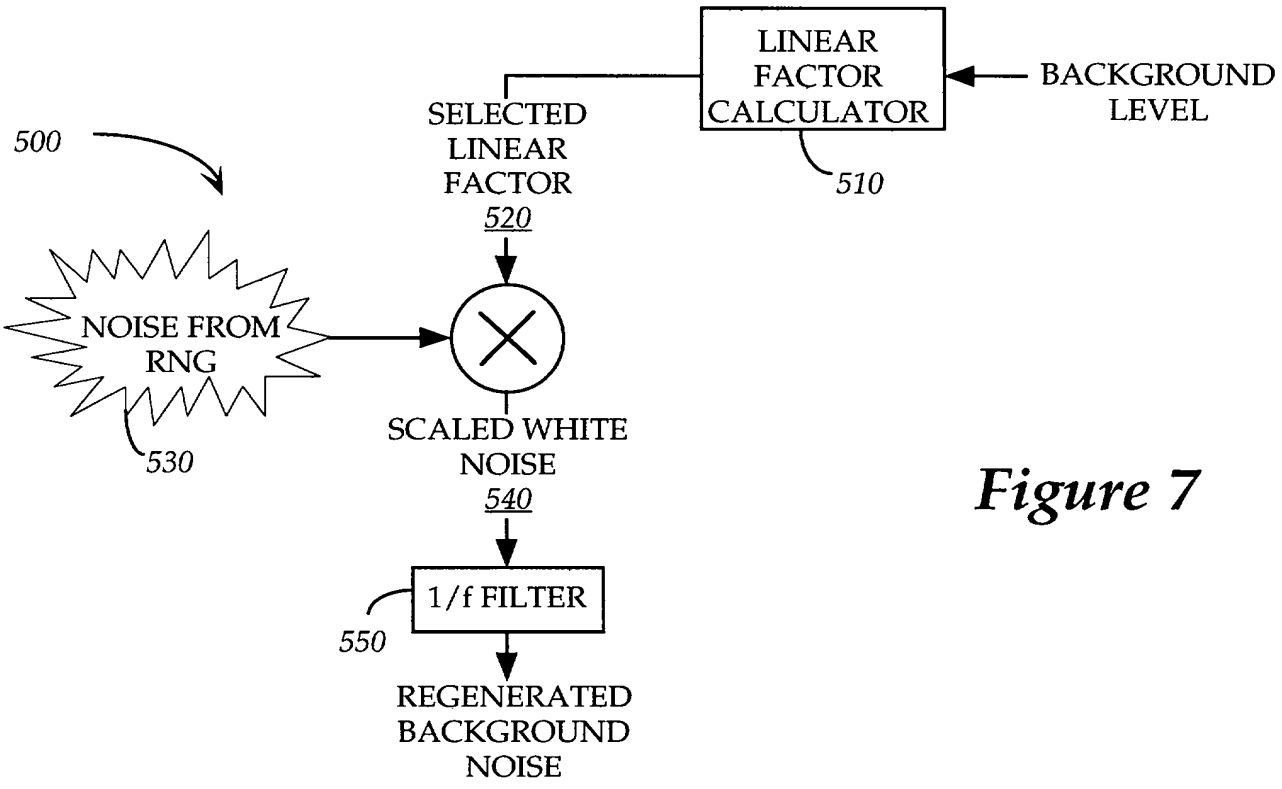

Figure 7

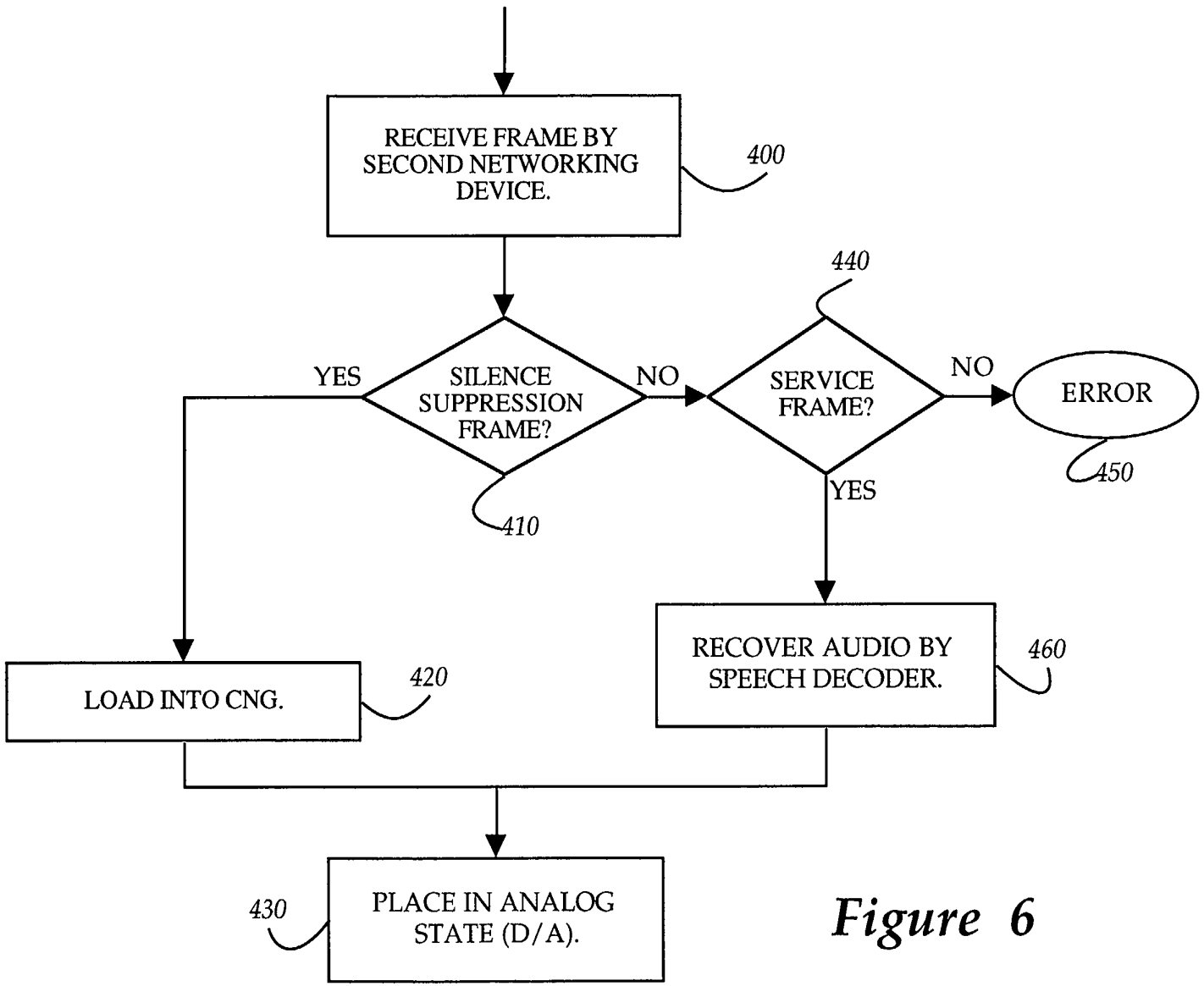
RECEIVE FRAME BY SECOND NETWORKING DEVICE.
400
SILENCE SUPPRESSION FRAME?
410
YES
NO
SERVICE FRAME?
440
NO
ERROR
450
YES
LOAD INTO CNG.
420
RECOVER AUDIO BY SPEECH DECODER.
460
PLACE IN ANALOG STATE (D/A).
430

Figure 6

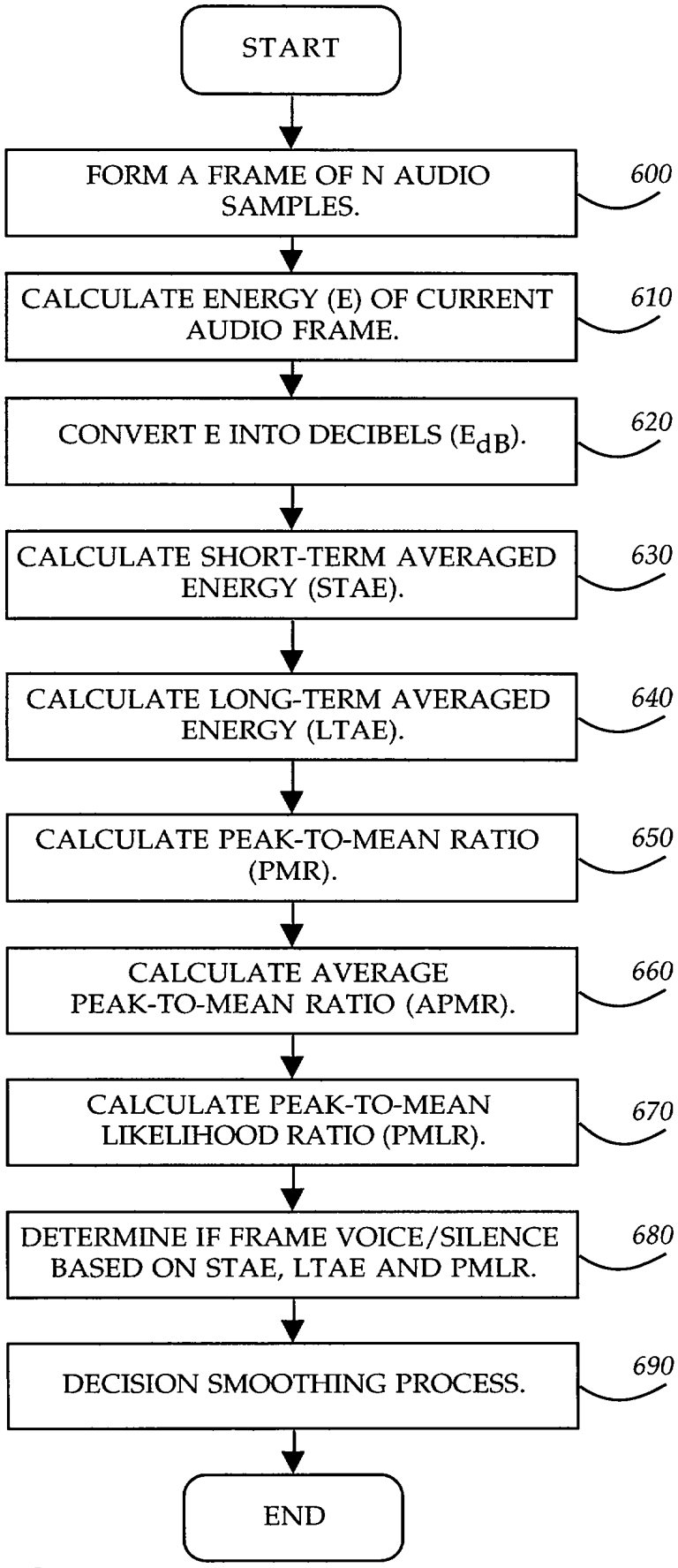
START
FORM A FRAME OF N AUDIO SAMPLES.
600
CALCULATE ENERGY (E) OF CURRENT AUDIO FRAME.
610
CONVERT E INTO DECIBELS (E_{dB}).
620
CALCULATE SHORT-TERM AVERAGED ENERGY (STAE).
630
CALCULATE LONG-TERM AVERAGED ENERGY (LTAE).
640
CALCULATE PEAK-TO-MEAN RATIO (PMR).
650
CALCULATE AVERAGE PEAK-TO-MEAN RATIO (APMR).
660
CALCULATE PEAK-TO-MEAN LIKELIHOOD RATIO (PMLR).
670
DETERMINE IF FRAME VOICE/SILENCE BASED ON STAE, LTAE AND PMLR.
680
DECISION SMOOTHING PROCESS.
690
END

Figure 8

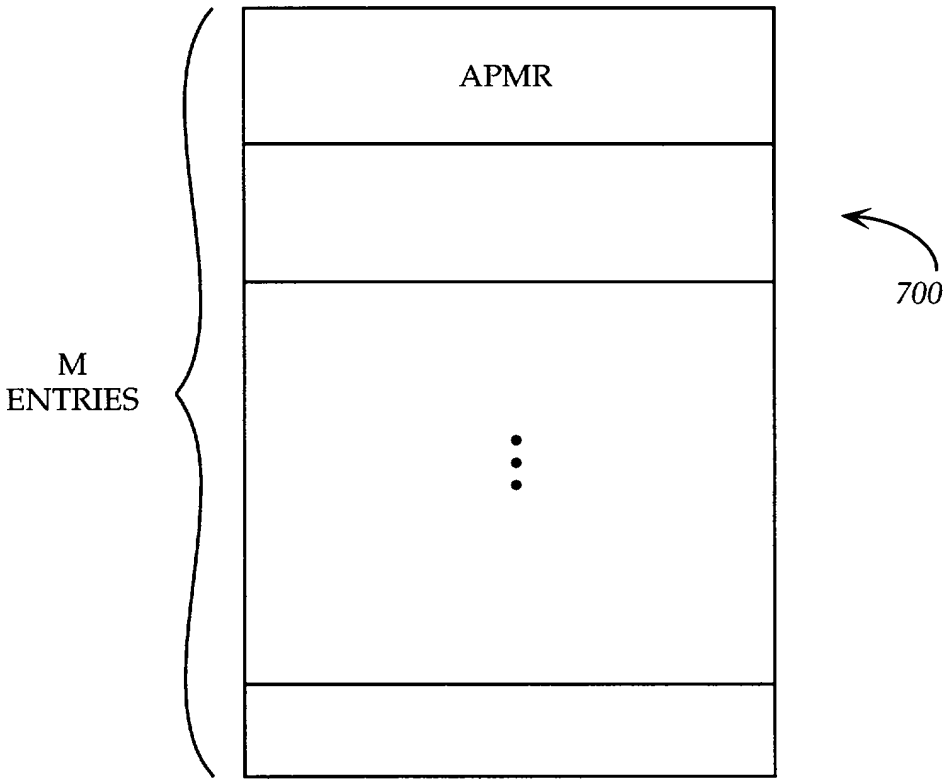

Figure 9

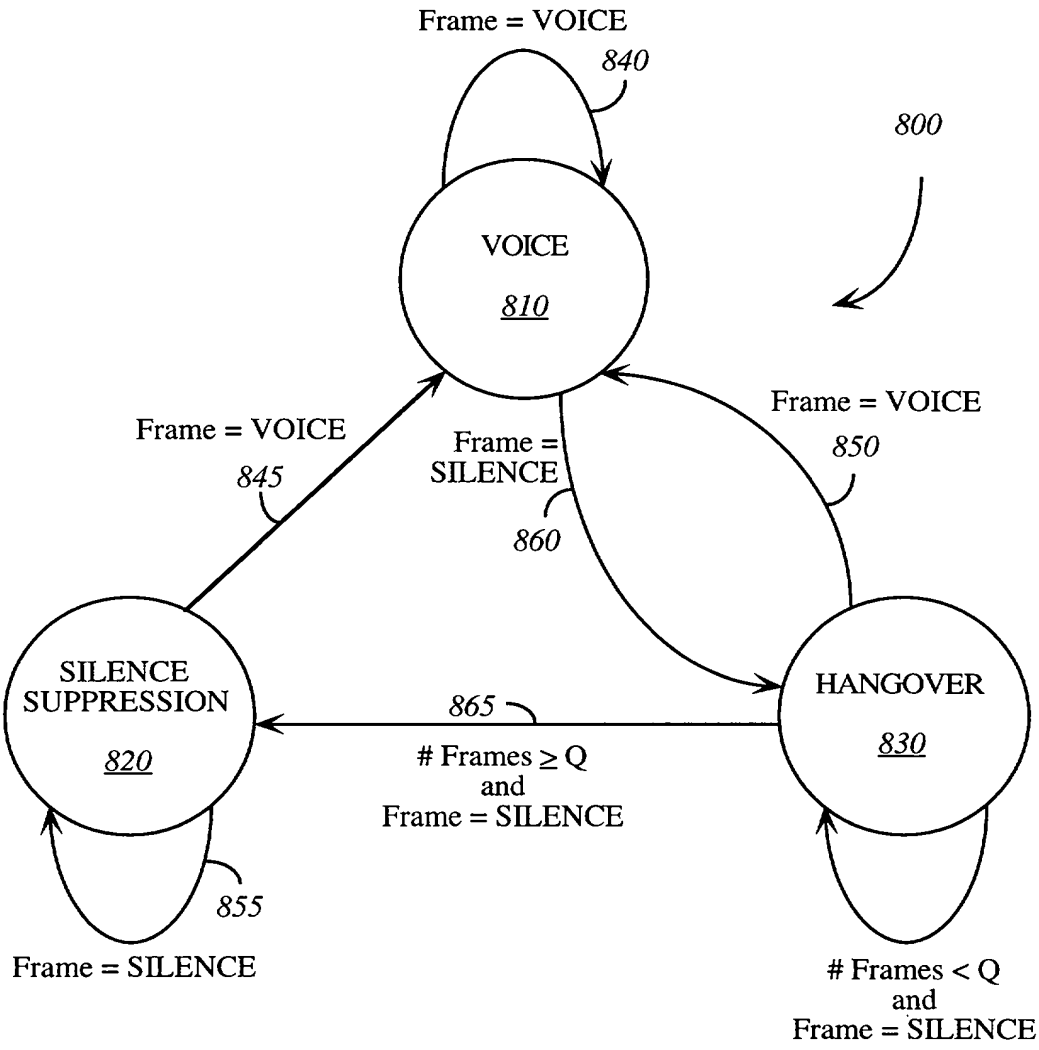
Frame = VOICE
840
800
VOICE
810
Frame = VOICE
845
Frame =
SILENCE
860
Frame = VOICE
850
SILENCE
SUPPRESSION
820
865
Frames ≥ Q
and
Frame = SILENCE
HANGOVER
830
855
Frame = SILENCE
# Frames < Q
and
Frame = SILENCE

Figure 10

Attorney's Docket No.: 003239.P010

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below, next to my name.

I believe I am the original, first, and sole inventor (if only one name is listed below) or any original, first, and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

A VOICE ACTIVITY DETECTOR FOR PACKET VOICE NETWORK

the specification of which [X] is attached hereto.
[ ] was filed on _____ as
United States Application Number _____
or PCT International Application Number _____
and was amended on _____ .
(if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claim(s), as amended by any amendment referred to above. I do not know and do not believe that the claimed invention was ever known or used in the United States of America before my invention thereof, or patented or described in any printed publication in any country before my invention thereof or more than one year prior to this application, that the same was not in public use or on sale in the United States of America more than one year prior to this application, and that the invention has not been patented or made the subject of an inventor's certificate issued before the date of this application in any country foreign to the United States of America on an application filed by me or my legal representatives or assigns more than twelve months (for a utility patent application) or six months (for a design patent application) prior to this application.

I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d), of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s):

APPLICATION NUMBER	COUNTRY (OR INDICATE IF PCT)	DATE OF FILING (day, month, year)	PRIORITY CLAIMED UNDER 37 USC 119
			☐ No ☐ Yes
			☐ No ☐ Yes
			☐ No ☐ Yes

I hereby claim the benefit under Title 35, United States Code, Section 119(e) of any United States provisional application(s) listed below:

APPLICATION NUMBER	FILING DATE

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

APPLICATION NUMBER	FILING DATE	STATUS (ISSUED, PENDING, ABANDONED)

I hereby appoint BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP, a firm including: Farzad E. Amini, Reg. No. P42,261; Aloysius T. C. AuYeung, Reg. No. 35,432; William Thomas Babbitt, Reg. No. 39,591; Carol F. Barry, P41,600; Jordan Michael Becker, Reg. No. 39,602; Bradley J. Bereznak, Reg. No. 33,474; Michael A. Bernadicou, Reg. No. 35,934; Roger W. Blakely, Jr., Reg. No. 25,831; Gregory D. Caldwell, Reg. No. 39,926; Kent M. Chen, Reg. No. 39,630; Lawrence M. Cho, Reg. No. 39,942; Yong S. Choi, Reg. No. P43,324; Thomas M. Coester, Reg. No. 39,637; Roland B. Cortes, Reg. No. 39,152; Barbara Bokanov Courtney, Reg. No. P42,442; William Donald Davis, Reg. No. 38,428; Michael Anthony DeSanctis, Reg. No. 39,957; Daniel M. De Vos, Reg. No. 37,813; Tarek N. Fahmi, Reg. No. P41,402; James Y. Go, Reg. No. 40,621; Richard Leon Gregory, Jr., P42,607; Dinu Gruia, Reg. No. P42,996; David R. Halvorson, Reg. No. 33,395; Thomas A. Hassing, Reg. No. 36,159; Phuong-Quan Hoang, P41,839; Willmore F. Holbrow III, Reg No. P41,845; George W Hoover II, Reg. No. 32,992; Eric S. Hyman, Reg. No. 30,139; Dag H. Johansen, Reg. No. 36,172; William W. Kidd, Reg. No. 31,772; Tim L. Kitchen, Reg. No. P41,900; Michael J. Mallie, Reg. No. 36,591; Paul A. Mendonsa P42,879; Darren J. Milliken, P42,004; Thinh V. Nguyen, Reg. No. P42,034; Kimberley G. Nobles, Reg. No. 38,255; Michael A. Proksch P43,021; Babak Redjaian, Reg. No. P42,096; James H. Salter, Reg. No. 35,668; William W. Schaal, Reg. No. 39,018; James C. Scheller, Reg. No. 31,195; Anand Sethuraman, Reg. No. P43,351; Charles E. Shemwell, Reg. No. 40,171; Maria McCormack Sobrino, Reg. No. 31,639; Stanley W. Sokoloff, Reg. No. 25,128; Allan T. Sponseller, Reg. No. 38,318; Steven R. Sponseller, Reg. No. 39,384; Geoffrey T. Staniford, P43,151; Judith A. Szepesi, Reg. No. 39,393; Vincent P. Tassinari, Reg. No. P42,179; Edwin H. Taylor, Reg. No. 25,129; George G. C. Tseng, Reg. No. 41,355; Lester J. Vincent, Reg. No. 31,460; John Patrick Ward, Reg. No. 40,216; Stephen Warfola, P43,237; Ben J. Yorks, Reg. No. 33,609; and Norman Zafman, Reg. No. 26,250; my attorneys; and Amy M. Armstrong, Reg. No. P42,265; Robert Andrew Diehl, Reg. No. P40,992; and Edwin A. Sloane, Reg. No. 34,728; my patent agents, with offices located at 12400 Wilshire Boulevard, 7th Floor, Los Angeles, California 90025, telephone (714) 557-3800, with full power of substitution and revocation, to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of Sole/First Inventor (given name, family name) Zifei Peter Wang

Inventor's Signature Zifei Peter Wang Date 8/14/98

Residence Thousand Oaks, California (City, State) Citizenship Peoples Republic of China (Country)

P. O. Address 535 N. Oak Tree Lane, #B

Thousand Oaks, California 91360 USA